ASCII article from DATAMATION

THE AMERICAN STANDARD CODE FOR INFORMATION INTERCHANGE

part one: review and preview

by R. W. BEMER, UNIVAC Division, Sperry Rand Corp., New York, N.Y.

It is very doubtful that Herman Hollerith ever considered, in 1905, that he would have to talk to Jean Baudot. After all, the man was dead. But this is exactly what is happening today in the inevitable marriage of computers and communications systems. The punched cards that Hollerith created must communicate with the punched paper tape of Baudot. The problem is that there is absolutely no logical similarity or relationship between the codes which represent the various letters, digits and other characters.

Hollerith designed his code for a mechanical counting reader. When cards became input to computers, as well as mechanical devices, a code correspondence had to be applied. In forming the IBM binary coded decimal (BCD) code, the 0 to 9 rows on the card were equated to 0000 through 1001, and thus the binary value corresponds to the decimal row value. Two more bits precede these to represent the four "zones" (12, 11, 0 and blank) by 00, 01, 10, and 11, although not respectively and indeed this varies among IBM equipment. Other manufacturers made different assignments in various attempts for internal economies. Assignments even vary among individual customers. Thus although most IBM users have the 12 punch as a plus sign and the 11 punch as a minus sign there are many others to whom the reverse is true.

There is a binary code inherent in the punched paper tape of Baudot, but this depends upon which tracks are made to correspond to which binary positions. Sorry to say, this choice has been made in several ways. Even so, Baudot did not make his assignment on a sequential basis for the digits or letters of the alphabet. Due to the technology of the time, it was done on the basis that the most frequently used characters would be represented by the fewest punched holes, to save wear and tear on the punch dies! To illustrate:

Letter	blank	E	T	A	O	I	N	S	H	R	D	L	U
No. of Punches	1	1	1	2	2	2	2	2	2	2	2	2	3

This should prove that there was never an actual person by that name! Apparently not much was known about digit frequency in those days, for they were assigned:

Digit	0	1	2	3	4	5	6	7	8	9
No. of Punches	3	4	3	1	2	1	3	3	2	2

Such technological conditions are largely removed now, and logical considerations assume commanding importance.

topsy in the information processing field

About four or five years ago many people awoke individually to the fact that we are in an almost impossible jumble in the coding of information. Consider the way it grew from the IBM standpoint, based upon the punched card. First came the digits 0 to 9, with & (or +) and −. So far there is only one problem, the duality of & and + as represented by the 12-punch. Now add zones for the letters. The digits 0 to 9 with a 12-punch mean A to I, with an 11-punch they mean J to R, and with a 0-punch

Mr. Bemer is an alternate member of X3.2, the X3.5 Glossary committee, and AFIP member of the IFIP TC-1 Terminology committee. Before joining Univac, where he is director of systems programming, he was with The RAND Corp. and IBM, and started the computing installation at Lockheed Missiles.

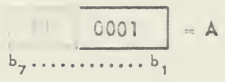
they mean / and S to Z. Simple. But now what? Without using the other combinations of two punches (e.g. 3-6) we move directly to combinations of three punches by adding 8's. This gives such characters as . , ° $ @ & and ◇. So far this can be lived with.

Now design the reader on the IBM 702 so that all illegal punch combinations are rejected; that is, only 48 out of the 4096 (2^{12}) possible combinations are legal. Now find out that more codes are needed for tape control in a computer, and so far only 48 out of a possible 64 codes available in a 6-bit character have been used. We try to see what happens for all 4096 combinations, and are surprised to find that the engineer goofed a little — nine supposedly illegal combinations slip by! So 0-2-8 is a record mark and 12-5-8 is a group mark.

On the other side of a high fence (between scientific and commercial computing at that time) the 704 people come up with FORTRAN, which needs the characters ( ) + and =, and certainly does not need % ◇ & and #. Since there are only 48 positions on the type wheels of the 407, a dual assignment is made. This makes it difficult for the installation with both scientific and commercial problems, but they learn to live with it. Along

Figure 1. *American Standard Code for Information Interchange*

Example:



0000	NULL	DC_0 ①	b	0	@	P	UNASSIGNED
0001	SOM	DC_1	!	1	A	Q	
0010	EOA	DC_2	"	2	B	R	
0011	EOM	DC_3	#	3	C	S	
0100	EOT	DC_4 (STOP)	$	4	D	T	
0101	WRU	ERR	%	5	E	U	
0110	RU	SYNC	&	6	F	V	
0111	BELL	LEM	'	7	G	W	
1000	FE_0	S_0	(	8	H	X	
1001	HT/SK	S_1	)	9	I	Y	
1010	LF	S_2	*	:	J	Z	
1011	V_{TAB}	S_3	+	;	K	[	
1100	FF	S_4	(comma)	<	L	\	ACK
1101	CR	S_5	-	=	M	]	②
1110	SO	S_6	.	>	N	↑	ESC
1111	SI	S_7	/	?	O	←	DEL

= 4 Bit Subset

comes the 1401 with its chain printer that has 240 character positions around it, normally in five sets of 48. But it could be in four sets of 60. Because it is to be a satellite machine for many large scale installations, the FORTRAN characters are given their own separate codes for programming convenience. Now we have both dual and individual assignments in the same installation. What confusion!*

Another trouble is that the internal bit assignment for the group mark is 11 1111, and when the code is filled out to the full 64 characters possible it is found that the counting mechanical reader demands that the punch combination for the group mark be 12-7-8, particularly for the 7070 and not 12-5-8 as for the 705. However, the same tape must be capable of being read by both machines.

And so it goes, each mistake by expediency being piled on top of the last one. And so many customers already use these incompatible devices that it just doesn't seem economical to change it now. Or could it be that things will get worse and we will wish we had straightened things out last year before it gets even more expensive?

It should not be thought that IBM is the only manufacturer with such problems. UNIVAC had a similar set of problems, particularly with both 80- and 90-column cards. The RCA 501 was designed with an internal code in which the letters and digits were assigned to consecutive binary numbers, a very sensible arrangement that makes data processing much easier. This is because there is something known as a "collating" or "ordering" sequence. If there were not, it would be very difficult to find a word in the dictionary or a number in the phone book. The 501 orders by simple binary comparison, with no extra hardware or wasted time. If this seems only reasonable, remember that IBM does not make any equipment with an internal code corresponding to its collating sequence, contrary to some beliefs. Ordering is done either by special hardware ($75 a month for early 1401's) or by programming, as on the 7090. Figure 2 shows two IBM cards, punched and interpreted. Columns 1 to 64 correspond to the binary sequence 00 0000 to 11 1111, or octal 00 to 77.

Let's see what happened to that 501. The 301 was designed to aim at the extensive punched card business. What could be more natural than to forget the 501 code and adopt the internal code of the IBM 704? Later a translator was built to convert codes in both directions between the 301 and 501. Just one problem, though — any file put in order on the 301 was out of order for the 501, and vice versa. At least without programming or additional hardware. This is hardly a trivial problem. IBM calculated in 1961 (in connection with ASA work) that it might take from $5,000,000 to $30,000,000 of machine time on the fastest computers just to reorder all existing files (as necessary - most would not require it, having only numeric keys) to a new collating sequence. This is the problem IBM faced in participating in code standardization work. If a standard code were to specify the collating sequence to be identical with the binary sequence, it would not match the IBM collating sequence. Makes even a big company stop and think; it might be hard to get the customer to take the broad view of future advantages and foot the bill.

However, occasions do arise when the situation is so muddled that desperate measures must be taken. As an example, Australia will change over to its new unit of currency, the "Royal", in February of 1966. This will replace the old pound and will be divisible into 100 cents,

(*Author's note—I'll take my share of the blame for some of this.)

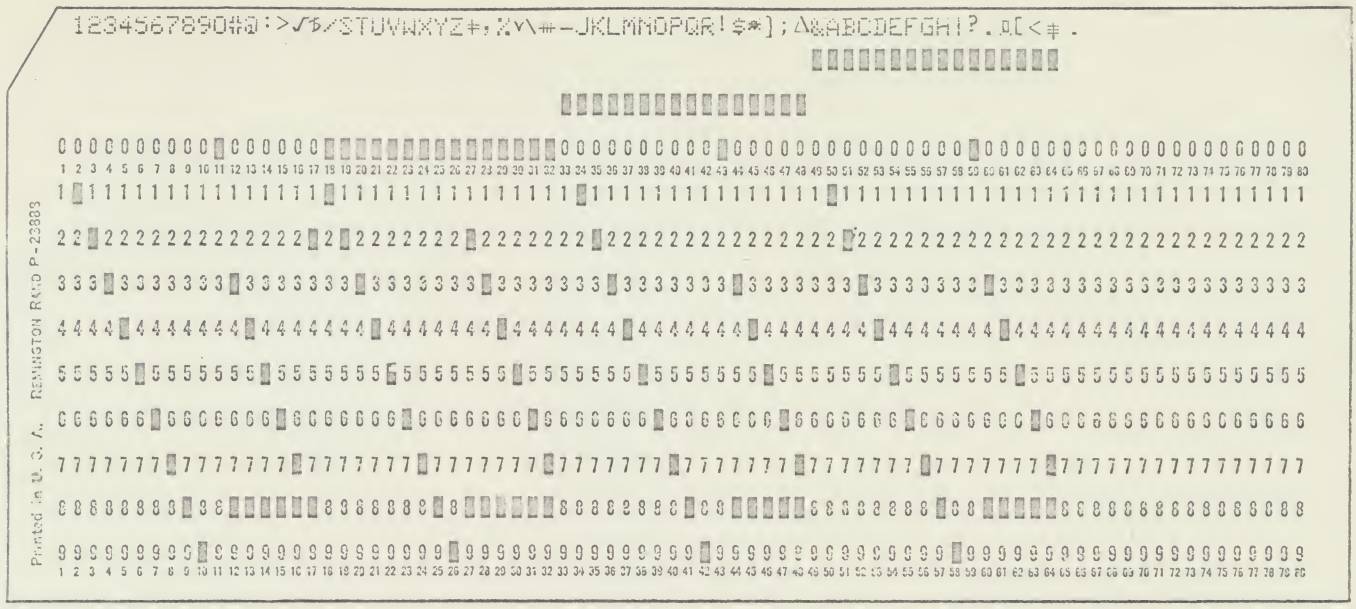


Figure 2

IBM STANDARD BCD INTERCHANGE CODE (1962)

just as the dollar is. In value it will be equivalent to 12 shillings of the pound system and about $1.12 US. Such a serious step will affect almost every area of the economy, from cash registers to coin changers, from education to counting procedures. The Aussies must think it worth it in the long run, however, and even England is now considering a change, perhaps to the decimal florin.

the american standard code

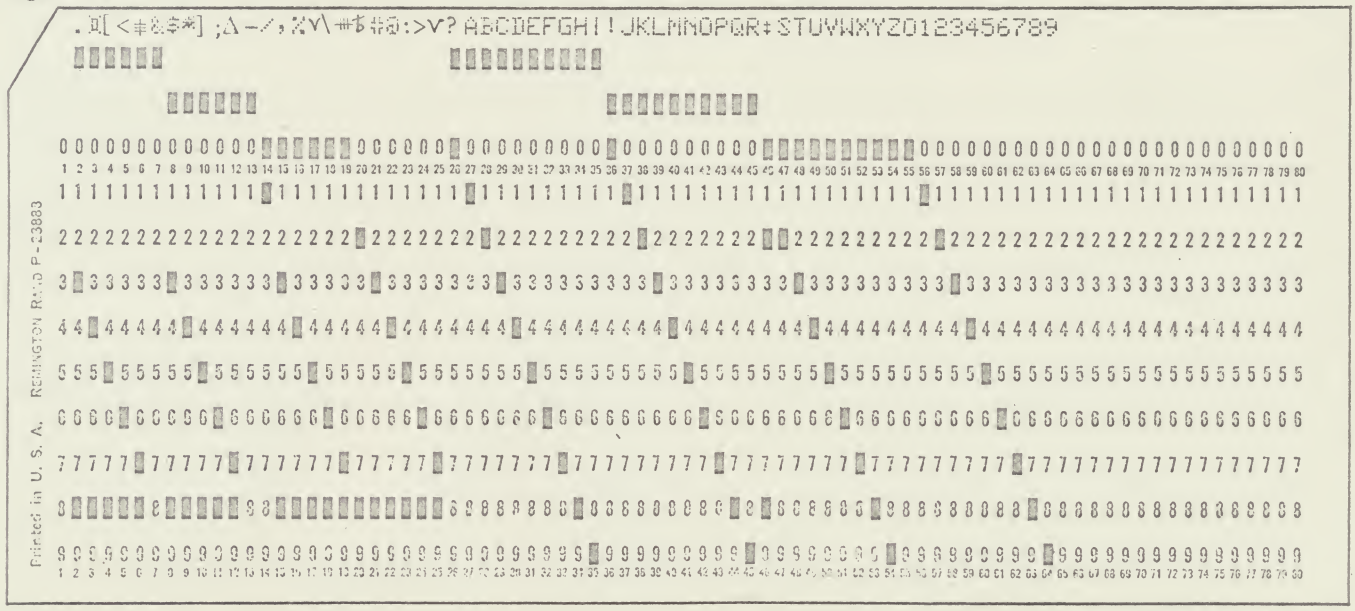
It was in a similar atmosphere of dissatisfaction with any existing system that the new American Standard Code for Information Interchange (ASCII) was developed, becoming an official ASA Standard No. X3.4 on 17 June 1963. The development and standardizing process was lengthy and sometimes turbulent. The important thing was that, as POGO says, "All was given equal chance to discuss and re-cuss." There was plenty of both. The problems of effective standardization are not new to readers of DATAMATION, but the successful adoption of a standard of this magnitude certainly is. Perhaps this success will help to reaffirm some faith in and support of these efforts.

The code was derived by the subcommittee method, one of the three ways by which an American Standard can be achieved, and certainly the most difficult. Whereas most standards are adoptions or reworks of existing practice, this code is a considerable departure from any previous code, although generic similarities to certain predecessors are certainly to be seen. Subcommittee X3.2 was chaired successively by representatives from IBM, Burroughs, and presently the Department of Defense, Navy Management Office. Several independent efforts were started in the 1958-59 period to deal afresh with the code problem. By universal agreement it was impossible to make enough sense out of existing codes; they just did not meet requirements evident at that time, nor did they provide for obvious future requirements. Among the major efforts were those of:

1. The Electronic Industries Association, a body which had produced many previous standards, working originally from a paper tape viewpoint, but later becoming general.
2. IBM, with the 8-bit code for STRETCH, which among other features provided for both upper and lower case alphabets.
3. The Department of Defense (Army Signal Corps)

Figure 2A

IBM STANDARD COLLATING SEQUENCE



which developed and sponsored the Fieldata code. Despite a few drawbacks, this was a great improvement upon existing codes and many of its features are to be seen in ASCII.

4. The British Standards Institution, which also started with the intention of standardizing paper tape codes and

The abbreviations used in Figure 1 mean:

NULL	Null/Idle	CR	Carriage return
SOM	Start of message	SO	Shift out
EOA	End of address	SI	Shift in
EOM	End of message	DC_0	Device control ① Reserved for Data Link Escape
EOT	End of transmission	DC_1-DC_3	Device control
WRU	"Who are you?"	ERR	Error
RU	"Are you . . . ?"	SYNC	Synchronous idle
BELL	Audible signal	LEM	Logical end of media
FE_0	Format effector	S_0-S_7	Separator (information)
HT	Horizontal tabulation	b	Word separator (blank, normally non-printing)
SK	Skip (punched card)	ACK	Acknowledge
LF	Line feed	②	Unassigned control
V/TAB	Vertical tabulation	ESC	Escape
FF	Form Feed	DEL	Delete/Idle

was drawn gradually into the whole data processing field.

5. The SHARE organization, which sought to coordinate their existing IBM equipment.

major features of the code

Let us examine the several salient and sometimes new features of the code and their significance:

1. As yet it is only a reference code. The particular representations in media such as punched cards, punched tape and magnetic tape are not yet defined, although they are perhaps implied in some respects.

2. It is (so far) a 7-bit code, with provision to expand to 8 bits as required. In the 7-bit form, a 6-bit subset of 64 codes is assigned completely to information characters, the other 64 so far are essentially control characters. This separation can be of convenience to equipment designers in the combined data processing and communications field; it also should produce many economies.

3. Although not yet stated in the standard, there is an implied collating sequence that may be used in the straight binary comparison mode. For IBM, which presently collates digits higher than the alphabet, an Exclusive OR device in passive logic can put the digit vector higher than the alphabet. This is no more than the existing device which allows the 709 family to write and read BCD tape.

4. The set can be collapsed in a regularized and prescribed manner, if required, into a 6-bit set for existing 6-bit machines and other equipments, to a 5-bit set for modification of existing Teletype and Telex sets (particularly in Europe), and even to a 4-bit set. This latter is of very special interest, in that it can be used for cash registers and other basically numeric-only devices, but at the same time may be used in the double numeric mode for computers internally (like the 650, 7070 and STRETCH). It is indicated by the offset shaded vector in the diagram of the code. The reason for the offset is that certain nondigit characters of the 4-bit set must collate lower than both digits and alphabet when appearing in ordering keys. The reverse expansion upward is a simple matter of passive logic.

5. Certain replacements (carefully checked out internationally) allow for non-American usage. For examples, the single digits 10 and 11 (sic) for English pence can replace the colon and semicolon in the digit vector, at least until they follow the lead of the Australians; the characters following the alphabet are of relatively low usage so that they may be replaced with the additional letters of expanded Roman alphabets, particularly as used by the Scandinavian countries.

6. The ESCape code (111 1110) provides for 127 alternate sets in the 7-bit set, 255 in the 8-bit set. Some of these sets may have official standing and some may be arbitrarily reserved to certain equipments. An example might be an alternate set with the Roman alphabet replaced by the Cyrillic alphabet, the unreplaced characters remaining unchanged. The ESCape character is usually followed by another code which is devoid of its usual meaning, by virtue of following the ESCape, and indicates which one of the alternate character sets is in force until the next ESCape character is encountered.

7. The two righthand vectors were purposely reserved as the logical places to put a lower case alphabet, if desired for this to be available in a single character mode. For lesser equipments, the upper case alphabet may be used in conjunction with the Shift In (000 1110) and Shift Out (000 1111) characters to produce a lower case facility.

8. Special consideration has been given for the characters required in programming and other special languages. All of the characters of the COBOL set are included. The ESCape characters may be used to shift to one or more special sets containing all of the characters of ALGOL (including the unique lower case alphabet). Other sets may be reserved for special languages of typesetting, information retrieval, graphic design, medical reports, etc. For example, a special set for numerical machine tool control could be an alternate 4-bit subset which is identical with the standard subset in the thirteen characters 0 - 9, decimal point, plus and minus; the other three characters would be replaced by X, Y and Z for axis symbols.

9. Note that many new characters have been introduced in the control area, particularly designed for self-delimiting of streams of characters. These may be hierarchic in nature, used to describe records, fields, subfields and so forth, or they may be syntactic in nature, indicating phrases, sentences, paragraphs, etc.

why a 7-level code?

Actually ASCII is an 8-level code with the eighth bit unassigned as yet. The new A. T. & T. system, supplied with terminal equipment by its subsidiary Teletype Corp., is based completely on eight bits between start - stop pulses. This is not only for future expansion but also for practical operations today. The eighth bit may be used for parity (preferably odd) if desired, and perhaps other uses may evolve. Basically, however, the 8-bit transmission unit was selected because eight is a magic number, being a power of two. In the information theory business there is nothing more economical than a power of two, and A. T. & T. knows it. Economy is important when you are creating a whole new system of this magnitude, and indeed that magnitude may be well up into the billions. There is even provision for an eleventh digit in the direct dialing system so Teletype can tell whether an 8-bit unit (Model 33) is talking to another 8-bit unit or to a 5-bit unit, or vice versa.

Several new computers are now being designed with 8-bit capabilities. At least one model, STRETCH, is in operation. Another is reported to use ASCII internally in

the double numeric mode. In certain 6-bit machines the word is designed to 48 bits to handle either six 8-bit characters or eight 6-bit characters. This code certainly facilitates transmission of pure binary data.

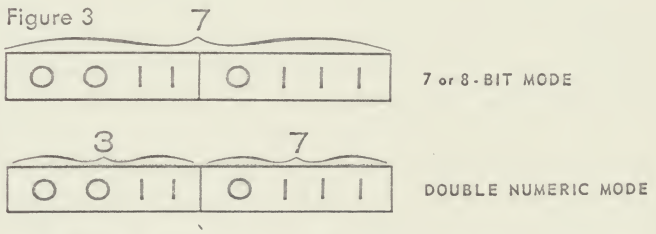
Subcommittee X3.2 appears to have no objection to the eventual assignment of meaning to all of the 256 codes in the 8-bit set. They sensibly avoided trying to be omniscient now and rather made adequate provision for expansion as further developments are made. Besides, they had to consider the Europeans and international standardization work in this area by ISO TC 97 on Computers and Information Processing. This work might not catch up to A. T. & T. for a while. Meanwhile the code will probably have to be adapted to 6-bit systems and even to five bits to work on existing Telex circuits, for the Europeans may not be able to install an entire new system in several countries in less than several years. The ASCII code is certainly set up to reduce the code size as required.

There are many advantages to having more unique codes in the set. There are some still unassigned in the 7-bit set, and of course nothing except a possible parity usage is assigned to the 8-bit set. There are several possible assignments for these spare codes, although none of these have been discussed extensively yet by X3.2:

1. A code which turns parity off and on, or possibly two individual codes, one for each of these two functions. This would facilitate compatibility between equipment using the 7-bit code with parity and other equipment desiring to utilize the full 8-bit set.

2. A code which says "repeat the transmission (it was bad) back to the last S_i code." Presumably this code would be followed by the particular S_i code required. This S_i would be sent back to the transmitting equipment, which would hold it in memory and search backwards along the transmitted stream until a match was found. The transmission would be restarted at that point, both sending and receiving equipment knowing exactly where to pick up again.

Figure 3

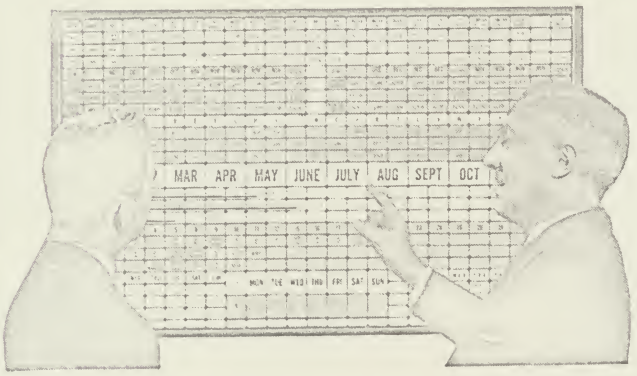


3. Codes to ignore normal communications control so that pure binary data may be transmitted without any character meaning. These will have to be handled carefully so that return to the normal transmission mode may be effected. This might have to be done by either timing the binary transmission, sending a predetermined number of 8-bit units with automatic return, or having the receiving device actuate the return through an extra channel.

4. Codes to switch to double numeric (two 4-bit digits within a single character) and back for reasons of economy of transmission in numeric only mode. ■

(Part two of Mr. Bemer's article will be published next month.)

You Get Things Done Better
By Seeing What's Happening



BOARDMASTER VISUAL CONTROL

★ Gives Graphic Picture of Your Operations in Color.
★ Facts at a Glance—Saves Time and Prevents Errors.
★ A Simple, Flexible Tool—Easily Adapted to Your Needs.
★ Easy to Use. Type or Write on Cards, Snap on Board.
★ Ideal for Production, Scheduling, Sales, Inventory, Etc.
★ Compact, Attractive. Made of Metal. 750,000 in Use.

Complete Price $49.50 Including Cards

FREE 24-Page ILLUSTRATED BOOKLET CG-20 Without Obligation

GRAPHIC SYSTEMS, Yanceyville, North Carolina

CIRCLE 25 ON READER CARD

THE CAT COMPUTER

Already enjoying wide acceptance without formal announcement is a multi-purpose on-line CAT Computer of Average Transients. One at Mayo Clinic has participated in a transatlantic experiment, averaging out brain wave signals transmitted from England via the Relay satellite. The results, interpreted and diagnosed, were then sent back by the same route. Present applications are in medical and clinical research.

The CAT 400B is a product of the Mnemotron Div. of Technical Measurement Corp., White Plains, N.Y. The count capacity of the memory is 100K per ordinate, with up to 400 ordinates. The CAT can sum and average responses from four varying inputs on-line, simultaneously calculating and displaying data on a built-in scope. The ability to store averages in successive quarters of memory makes it possible to compare successive runs of averages without interrupting experiments (there is no theoretical limit to the number of responses which may be summed). Averages may be displayed on the 3″ CRT and on an X-Y Plotter Readout. The computer measures about one cubic foot, weighs 38 pounds, and consumes 30 watts. Price is $12K. ■